Title	REGRESSION ANALYSIS AND ERRORS IN VARIABLES IN ECONOMIC TIME SERIES
Author(s)	Abe, Osamu
Citation	Kyoto University Economic Review (1956), 26(2): 39-46
Issue Date	1956-10
URL	http://hdl.handle.net/2433/125426
Right	
Type	Departmental Bulletin Paper
Textversion	publisher

VOLUME XXVI NUMBER 2

Kyoto University Economic Review

MEMOIRS OF THE FACULTY OF ECONOMICS
IN THE KYOTO UNIVERSITY

OCTOBER • 1956

PUBLISHED BY THE FACULTY OF ECONOMICS
KYOTO UNIVERSITY • KYOTO, JAPAN

REGRESSION ANALYSIS AND ERRORS IN VARIABLES IN ECONOMIC TIME SERIES

By Mr. Osamu ABE*

(1)

In putting to the test the premises of the wellknown theory of stability of exchange, "whether the depreciation is effective for the recovery of equilibrium of the balance of trade", many studies contend that we cannot expect too much from the effectiveness of the lowering of the rate of exchange because the price elasticities of the import demands of various countries are far smaller than unity. But in most cases there seems to be cherished some hope behind it that "this is a short run analysis. Therefore, in the long run, there will be a certain improvement in the balance of trade as a result of the work of income effect and other forces"[1]. On the other hand, when the depreciation is discussed as an actual policy, it is often said that "the lowering of exchange rate may temporarily be effective, but its effect will soon be offset by the domestic inflation", referring to the recent case in Mexico. It is very interesting to see that this difference between these two views is relevant to the opposition of the classical and the modern schools in the theory of international trade[2], and in this case the thing approved by both two standpoints as one of their concession is the disproof of Orcutt that the measurement of price elasticities and the statistical data are inappropriate and unreliable, and consequently the result of calculations tends toward zero[3]. This disproof is a matter touched with the intrinsic nature of how to deal with the statistical errors in the regression analysis in economic time series on the whole and therefore, will be discussed systematically in this essay.

---

* Assistant Professor of Economics and Statistics, Tokyo Institute of Technology.

1) e. g, L. Metzler, "The Theory of International Trade," *A Survey of Contemporary Economscs*, Vol. 1. H. S. Ellis, ed. 1948, p. 245.

2) See J. Viner's view in the introductions of *International Economis—Studies*, 1951, and *International Trade and Economic Development*, 1953 together with the rebutment of G. Haberler, "The Relevance of the Classical Theory under Modern Conditions", *American Economic Review*, May 1954.

3) G. H. Orcutt, "Measurement of Price Elasticities in International Trade," *Review of Economics and Statistics*, May 1950.

(2)

Let us draw an instance at first. With an intention of testing statistically the applicability of the theory of comparative costs, MacDougall tried to examine "whether the United States (the United Kingdom) has a comparative advantage in those commodities that the United States (the United Kingdom) exports more than the United Kingdom (the United States)" in a form that "whether the relative price is cheaper in the United States (the United Kingdom)" in the first place[4]. For this purpose, he chose 109 items of those products exported from the both countries and calculated on the basis of the data between 1934 and 1938 the relative quantities of exports or $\frac{\text{Total Quantity of U. S. exports, 1934-38}}{\text{Totol Quantity of U. K. exports, 1934-38}}$ and the relative prices or $\frac{\text{Average Value of U. S. exports, 1934-38}}{\text{Average Value of U. K. exports, 1934-38}}$ for each product, from which he obtained the regression line shown in fig. 1.[5] The correlation coefficient is −0.74 and the slope is −3.6. That is, according to the data "for the period 1934-38, a difference of 1% in relative price tended to be associated with a difference of 3.6% in relative quantity of exports", but MacDougall discusses the possibility of a considerably flatter slope of this regression line, and points out the following grouds of his argument, contending that if there is no error in the statistical data, a difference of 1% in relative price must tend to be associated with a difference of at least 4–4.5% in relative quantity of exports.

When we disregard the transport costs and suppose there exists a perfect competition in the international market, and the theory of comparative costs applies perfectly, the relationship between relative price and relative quantity of exports of both countries must look like fig. 2, if there is no error in the statistical data. In other words, where America's price is less than the British even a bit, Britain would export nothing, and where Britain's price is less than the American, America would export nothig. But even in this case, if the statistical data includes some errors of observation, the observed points would then be more or less scattered as are shown in fig. 3, and the regression line which minimises the sum of the squares of the deviations in a horizontal direction would necessarily tend to slope. In other words, when independent variable in the regression equation is associated with errors of observation and the estimate of regression coefficient is calculated in an ordinary way, the value usually tends to be smaller

4) G. D. A. MacDougall, "British and American Exports: A Study Suggested by the Theory of Comparative Costs," *Economic Journal*, Dec. 1951.

5) Logarithmic graduation is used in the figure.

Fig. 1
Relative Prices & Quantities of U. S. & U. K. Exports of 109 Manufactures 1934–38

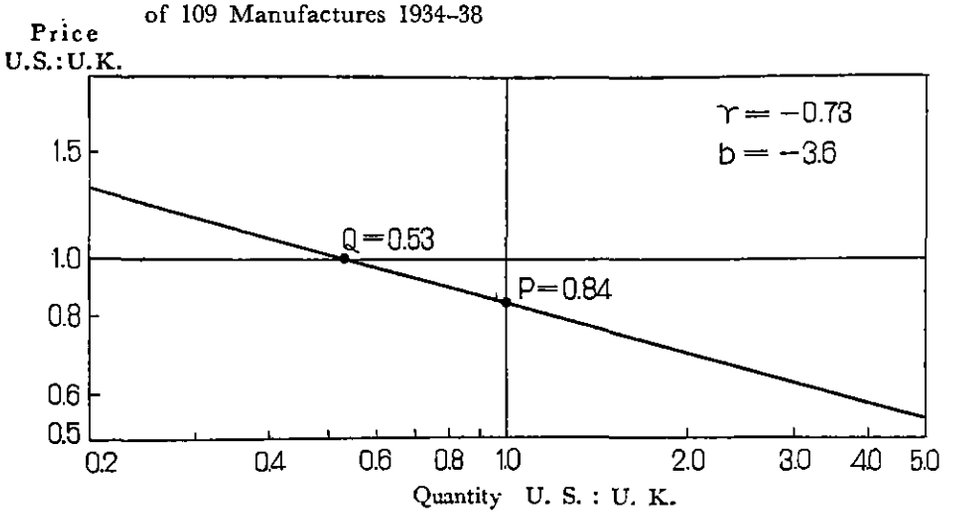


Fig. 2

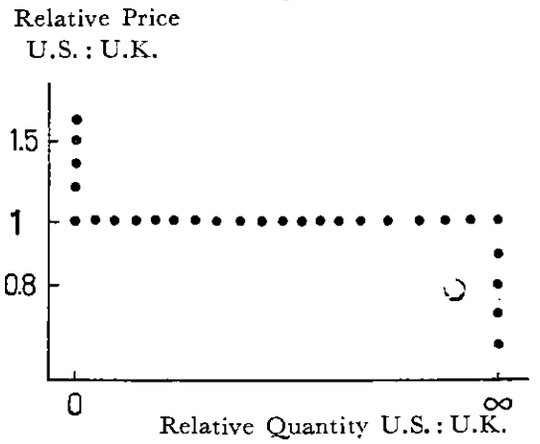
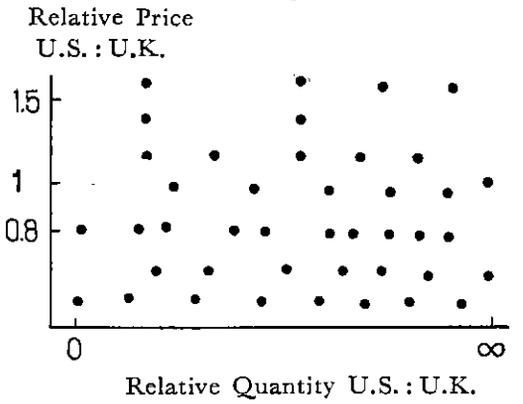


Fig. 3



than the true regression coefficient.

If this relationship is shown algebraically, it may look like this. Now, let X's be the deviations from the mean of the observed values of relative prices, and $Y's$ be that of the relative quantities of exports, and let each deviation be the sum of the true value (x, y) and the error of observation (ξ, η) respectively, then we have

$$X=x+\xi \qquad Y=y+\eta \quad . \tag{1}$$

Assuming that

$$\begin{aligned} E(X)&=E(x)=0 \\ E(Y)&=E(y)=0, \end{aligned} \tag{2}$$

true regression coefficient can be given by

$$=\frac{E(xy)}{E(x^2)} \quad . \tag{3}$$

But the expected value of the empirical regression coefficients which are given by the least squares method from the obseved values $X's$ and $Y's$ is approximately like this.

$$E(b)\sim \frac{E(XY)}{E(X^2)}=\frac{E(x+\xi)(y+\eta)}{E(x+\xi)^2}$$
$$=\frac{E(xy+x\eta+y\xi+\xi\eta)}{E(x^2+2x\xi+\xi^2)} \tag{4}$$

We assume

(i) no correlation between the true value and the error of observation of each variable, that is

$$E(x\xi)=E(y\eta)=0,$$

(ii) no correlation between the error of observation of dependent variable and the true value of independent variable, that is

$$E(x\eta)=E(y\xi)=0,$$

(iii) no correlation between the errors of observation of both variables, that is

$$E(\eta\xi)=0$$

and hence, from (4)

$$E(b)\sim \frac{E(xy)}{E(\lambda^2)+E(\xi^2)}=\frac{1}{1+\lambda^2}\beta, \tag{5}$$

$$\text{here} \qquad \lambda^2=\frac{E(\xi^2)}{E(x^2)}\geqq 0.$$

Therefore, $E(b)\leqq\beta$. Thus, MacDougall's argument was verified.

(3)

Now, the matter in question is that when the statistical data contain some errors how should we deal with the bias of the estimate of regression coefficient described above. This question has already become classic to a certain extent and we see many devices toward its solution[6]. But it seems to me that all of them left many difficult points in their application to economic time series. Besides, in empirical study the question is apt to be thoroughly neglected and mere calculation of classical regression coefficient is deemed satisfactory. But in this case the presumption that the estimated of regression coefficient is the unbiased one must be implicitly accepted. It is extremely doubtful whether this kind of presumption as it is be admissible in economic phenomena.

Now, let formularize again the whereabout of the question in conformity with the following argument. For simplicity sake, we shall consider

6) e. g, T. C. Koopmans, *Linear Regression Analysis of Economic Time Series*, 1937

the regressive relationship between the two variables X and Y, and let constant terms be zero. In addition to errors of observation of variables, we introduce the equational error ϵ which represents the disturbance of the dependent variable. Whence, we have

$$y_i=\alpha x_i+\epsilon_i \quad (i=1,\ 2,\cdots\cdots,\ n)$$
$$Y_i=y_i+\eta_i \qquad X_i=x_i+\xi_i$$

Here, x and y represent the true values of variables and ξ and η the errors of observation respectively.

Now, we shall frame the following hypotheses in regard ξ, η, and ϵ.

(i) $E(\eta_i \mid x_1,\cdots\cdots,x_n\ ;\ \xi_1,\cdots\cdots,\ \xi_n)=0$

(ii) $E(\xi_i \mid x_1,\cdots\cdots,\ x_n)=0$

(iii) $E(\epsilon_i \mid x_1,\cdots\cdots,\ x_n\ ;\ \xi_1,\cdots\cdots,\ \xi_n)=0$

($i=1,\ 2,\cdots\cdots,\ n$ in all cases)

(i) shows that the error of observation of dependent variabie is independent either of the true value of, or of the error of observation of indendent variable. (ii) shows that there is no correlation between the true value of and the error of observation of independent variable, and (iii) specifies that the equational error has no correlation with independent variable. In this case, we assume that the means of probability distribution of each error are all zero[7]

Now, when we seek for the expected value of the estimates of β by the least squares method, we can easly obtatn as shown in the previous section

$$E(b)\sim \frac{E(xy)}{E(x^2)+E(\xi^2)}$$
$$=\frac{1}{1+\dfrac{E(\xi^2)}{E(x^2)}}\beta \qquad (6)$$

To deal with this bias, between $E(b)$ and β, Berkson has recently developed an ingenious device.[8] According to him, X rather than x can be regarded as the predeteraming variable in many statistical data, particularly those of laboratory investigation of natural science. In these cases, it is better to regard the observed value X rather than the true value x as a constant. Considering that x would rather vary due to the existence of the error ξ, we replace the hypothesis (ii) referred above with the hypothesis

(iv) $E(\xi_i \mid X_1,\cdots\cdots X_n\ ;\ x_1,\cdots\cdots,\ x_n)=0$

---

7) In time series there may arise the problem of self-correlation error beside this, but we do not need at present any assumption regarding with it.

8) J. Berkson, "Are There Two Regressions?" *Journal of American Statistical Association,* Jume 1950

$$(i=1,\ 2,\cdots\cdots,\ n)$$

and we have

$$\frac{E(\xi^2)}{E(x^2)}=\frac{E\xi(X-x)}{E(x^2)}=0\,,$$

whence, from (6)

$$E(b)=\beta.$$

In other words, b becomes the unbiased estimate of β. Further it is proved that when we assume ξ, η and ϵ has a nomal distribution and no serial correlation respeetively the Fisherian reliability test is also applicable.

The device of Berkson described above cleverly attacks the points hitherto overlooked, but I wonder whether it is at all appropriate to assume this kind of errors in economic statistical data. For instance, the following example may come across our minds. Considering that capital coefficient is an economic varable and lacking in statistical data regarding the national capital necessary for the determination of its value, we assume from the actua lresults of other countries that our is also 4 and that the true value fluctuates around 4 every year. In this case, if we premise that the fluctuation behavior has a symmetric distribution, centering in 4, the hypothesis (iv) may work. But this sort of example is quite rare, and errors of observation concerining the bulk of economic data seem to have a characteristic of lying between the above mentioned hypotheses (ii) and (iv).

But even in this case we can resort to a fairly effective method to obtain the unbiased estimate of regression coefficient after the model of the device of Berkson. That is the grouping of independent variables. In imitation of Berkson's device mentioned above, we divide into several sections the range of variation of the true value x and make the means of these sections represent x. If we assume in this case that x contains no error and that in each section x distributes symmetrically, centering in respective representation values, it may be considered that errors between each section due to grouping would satisfy thc hypothesis (iv) with its expected value of zero. Therefore,

$$b=\frac{\Sigma XY}{\Sigma X^2}$$

will be the unbiased estimate of β.

Now the datum we can deal with is the observed value x attended with errors ξ. Assuming that grouping is not effected by these errors, the mean of each group is independent of errors of observation. Therefore, if error of observation has no correlation with the true value x and the mean is zero, we can premise the hypothesis (iv) by taking ξ as the compound of error of observation and the error caused by the representation of sections

divided by grouping by the means. Consequently, we will be able to obtain the unbiased estimate of regression coefficient provided the foregoing conditions are satisfied precisely. And even if those conditions which specifies the hypothesis (iv) may not be precisely satisfied, it still seems possible at least to lessen the bias of b by grouping.[9] The foregoing argument is not limited to the case of two variables, but may be extended to the cases of multiple regression in general.

(4)

It is known by the recent studies of Reiersøl and Geary that when the error of observation has such a character as specified by the hypothesis (i) and (ii), we can use instrumental variable as a means of removing or lessening the bias of the estimate of regression coefficient.[10] In this section, we shall consider such devices.

Let us assume a certain variable z which has correlation with the true value of the independent variable x, but is independent of the errors ξ or ϵ. A variable like z is called an instrumental variable. If we put

$$b'=\frac{\Sigma zY}{\Sigma zX}=\frac{\Sigma z(\beta x+\epsilon)}{\Sigma z(x+\xi)}$$

we have

$$E(b')\sim\frac{\beta E(zx)}{E(zx)}=\beta$$

and therefore b' can be the unbiased estimate of β.

But, in this case the variance of b' or $V(b')$ is generally larger than that of b, $V(b)$, and therefore, to use b' as the estimate makes less degree of precision. If so, it may be better for us to be satisfied with "biased estimate" to be obtained by the ordinary least squares method rather than to take the trouble of employing an instrumental variable to obtain an unbiased estimate. If we assume $\xi=\eta=0$, we can easily see the variance of the difference between b and b' depends upon the variance ratio of the two.

Now, the things which comes into question when we adopt the instrumental variable method is how to find such a variable that is highly

---

9) We may avoid the bias to be brought in by grouping errors by taking the means for representation values between sections, but we cannot avoid the bias caused by the erros of observation of variables.

10) O. Reiersøl, "Confluence Analysis of Lag Moments and Other Methods of Confluence Analysis," *Econometrica*, Jan. 1941. Ditto, "Identifiability of Linear Relation between Variables which are Subject to Error, "*Econometrica*, July 1950. Ditto, *Confluence Analysis by means of Instrumental Sets of Variables*, 1945. R. C. Geary, "Determination of Linear Relations between Systematic Parts of Variables with Errors of Observation the Variances of Which are Unknown," *Enonometrica*, Jan.1949.

correlated with x and yet is independedt of ξ. We can find various devices for this, too. For instance, it may be a way to do $z=+1$ or 0 or $1-$ (for $x \gtreqless$ its median), as Wald has informedly done.[11]

It may be also a good idea to establish ranking according to the size of X and to take ranking number as z. In either case, they will be efficient methods provided that the error of observation ξ is not very large. In case the value of ξ is pretty large and is supposed to exert a considrable influence upon the ranking of X, we may set up ranking to means X for every section by the grouping method described in the preceding section.[12]

It is wellknown that with the development of model analysis the so-called simultaneous equation approach has been adopted for measuring economic relation in lieu of the least squares method analysis by single equation which had formerly been used.[13] In the end we shall briefly touch up on the relation between the simultaneous equation approach and the instrnmental variable method.

When we set up a simultaneous equation of the regression equation of y' and x' and that of x' and instumeltal variable z, we have

$$\begin{cases} y=\beta x+\epsilon \\ x=\gamma z+\kappa \end{cases}$$

whence,

$$\begin{cases} Y=\beta X+\xi \\ X=\gamma z+\eta \end{cases} \qquad \text{provided} \begin{cases} \xi=\epsilon-\beta\xi \\ \eta=\kappa+\xi \end{cases}$$

Now, assuming that the error terms ξ and η are independent of z and their means are zero, we obtain as shown below the estimate of β by the reduced method so far as γ is not zero.

$$b'=\frac{\Sigma XY}{\Sigma Xz}$$

In this case, if ϵ and κ has a jointly normal distribution, b' would become the maximum likelihood estimate of β. And also it has already been demonstrated by Anderson and Rubin that if some qualifications are made as to the nature of jointly distribution, b' would become the undiased estimate.[14]

---

11) A. Wald, "The Fitting of Straight Lines if Both Variables are Subject to Error." *Annals of Mathematical Statistics*, No. 3, 1940

12) In his papers of 1941 referred above, Reiersøl uses the lagged values of x as an instrumental variable where x has serial correlation and ξ has not.

13) T. C. Koopmans and Others. "Statistical Inference in Dynamic Economic Models," 1950

14) T. W. Anderson and H. Rubin, "The Asymptotic Properties of Estimates of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *Annals of Mathematical Statistics*, March 1950.